

明 細 書

テキスト処理方法／プログラム／プログラム記録媒体／装置

技術分野

[0001] 本発明は、文字列や単語列といったテキスト文書を、意味的にまとまった部分ごとに、すなわち話題ごとに分割するテキスト処理方法／プログラム／プログラム記録媒体／装置に関する。

背景技術

[0002] この種のテキスト処理方法／プログラム／プログラム記録媒体／装置は、長大かつ多数のテキスト文書を意味内容ごとに、すなわち話題ごとに分割、分類等することによって、人がテキスト文書から所望の情報を得やすいように加工することを目的として用いられている。ここでテキスト文書とは、例えば、磁気ディスク等の記録媒体に記録された任意の文字や単語などの並びである。あるいは、紙に印刷されたり、タブレットに手書きされたりした文字列を光学的読取り装置(OCR)で読み取った結果や、人の発話で生じる音声波形信号を音声認識装置で認識した結果等も、テキスト文書である。さらに一般的には、毎日の天候の記録、店舗における商品の販売記録、コンピュータを操作した際のコマンドの記録、等々、時系列的に生成される記号の並びのほとんどは、テキスト文書の範疇に入る。

[0003] この種のテキスト処理方法／プログラム／プログラム記録媒体／装置に関して、大別して2種類の従来技術が挙げられる。これら2種類の従来技術について、図面を参照して詳細に説明する。

[0004] 第1の従来技術は、入力テキストを単語の系列 o_1, o_2, \dots, o_T として、系列中の各区間で単語の出現傾向に関する統計量を算出し、この統計量に急激な変化がみられる位置を話題の変化点として検出する。例えば図5に示すように、入力テキストの各部分に対して一定幅の窓を設定し、窓内における単語の出現回数を計数し、単語の出現頻度を多項分布の形式で算出する。そして、近接する2つの窓(図5における窓1および窓2)の間の差異が所定のしきい値より大きければ、これら2つの窓の境界で話題の変化が起こったと判定する。2窓間の差異には、例えば式(1)で表されるような、窓

ごとに計算された多項分布間のKLダイバージェンスを用いることができる。

[0005] [数1]

$$\sum_{i=1}^L a_i \log \frac{a_i}{b_i} \quad \cdots (1)$$

[0006] ここで、 a_i, b_i ($i=1, \dots, L$) はそれぞれ窓1、窓2に対応する単語の出現頻度を表す多項分布で、 $a_1 + a_2 + \dots + a_L = 1$, $b_1 + b_2 + \dots + b_L = 1$ を満たす。Lは入力テキストの語彙数である。

[0007] 上では特に、窓内の統計量を個々の単語の出現頻度から計算する、いわゆるユニグラム(unigram)としているが、隣接2つ組、3つ組、さらには任意個の組の単語出現頻度(それぞれバイグラムbigram、トライグラムtrigram、n-gram)を考えてもよい。あるいは、「2001年11月、情報処理学会論文誌、第42巻、第11号、第2650～2662頁、別所克人、単語の概念ベクトルを用いたテキストセグメンテーション」(文献1)に記載されているように、隣接しない単語同士の共起(すなわち、隣接しない複数の単語が同一の窓内に同時に出現すること)を考慮することにより、入力テキスト中の各単語をベクトルに置き換えて、このベクトルの移動量の多さで話題の変化点を検出することもできる。

[0008] 第2の従来技術は、種々の話題に関する統計的モデルをあらかじめ準備しておき、それらのモデルと入力単語列の最適なマッチングを計算することにより、話題の推移を求める。第2の従来技術の例は、「2000年、プローシーディング・オブ・フォース・ユーロピアン・カンファレンス・オン・リサーチ・アンド・アドバンスト・テクノロジ・フォー・デジタル・ライブラリ、アマラル他、トピック・ディテクション・イン・レッド・ドキュメント(Amaral et al., Topic Detection in Read Documents, Proceedings of 4th European Conference on Research and Advanced Technology for Digital Libraries, 2000)」(文献2)に記載されている。この第2の従来技術の例は、図6に示すように、「政治」、「スポーツ」、「経済」などといった話題ごとに、話題ごとの統計モデル、つまり話題モデルを作成して準備しておく。話題モデルは、あらかじめ話題ごとに大量収集されたテキスト文書から求めた単語出現頻度(ユニグラム、バイグラム等)である。このように話題モデルを準備し、これら話題間の遷移の起こりやすさ(遷移確率)を適宜決めておけ

ば、入力単語系列ともっともよく整合する話題モデル系列を機械的に算出することができる。仮に、入力単語系列を入力音声波形と置き換えて、話題モデルを音素モデルに置き換えてみれば容易にわかるように、音声認識に関して多数ある従来技術と同様に、DPマッチングの要領で、フレーム同期ビームサーチなどの計算法を利用して話題の遷移系列を計算することができる。

- [0009] 上で述べた第2の従来技術の例は、「政治」、「スポーツ」、「経済」など、人間が直感的に理解しやすい話題を設定して、話題の統計モデルを作成しているが、「1998年、プロシーディング・オブ・インターナショナル・カンファレンス・オン・アコースティック・スピーチ・アンド・シグナル・プロセッシング98、第1巻、333～336頁、ヤムロン他、ヒドゥン・マルコフ・モデル・アプローチ・トゥ・テキスト・セグメンテーション・アンド・イベント・トラッキング (Yamron et al., Hidden Markov model approach to text Segmentation and event tracking, Proceedings of International Conference on Acoustic, Speech and Signal Processing 98, Vol.1, pp.333–336, 1998)」(文献3)に記載があるように、テキスト文書に対して何らかの自動クラスタリング手法を適用して、人間の直感とは無関係な話題モデルを作る例もある。この場合、話題モデルを作るために大量のテキスト文書を話題ごとに分類しておく必要がないので、手間は幾分少なくてすむ。ただし、大規模なテキスト文書集合を用意して、そこから話題モデルを作成するという点は同様である。

発明の開示

発明が解決しようとする課題

- [0010] しかしながら、上述した第1の従来技術および第2の従来技術は、それぞれいくつかの問題を有する。
- [0011] 第1の従来技術では、窓間の差異に関するしきい値や、単語出現回数の計数範囲を規定する窓幅といったパラメータを最適に調整することが難しいという問題がある。あるテキスト文書に対して所望の分割がなされるようにパラメータ値を調整することは、可能な場合もある。しかし、そのために試行錯誤的にパラメータ値を調整する手間が必要である。加えて、仮にあるテキスト文書に対して所望の動作が実現できたとしても、同じパラメータ値を別のテキスト文書に適用した場合、期待通りに動作しないこ

とが多い。なぜなら、例えば窓幅というパラメータは、大きくすればするほど窓内の単語出現頻度を正確に見積もることができるから、テキストの分割処理も正確に実行できるが、窓幅は入力テキスト中の話題の長さよりも長いと、明らかに話題分割という当初の目的を達せられなくなる。すなわち、入力テキストの性質によって、窓幅の最適値は異なる。窓間の差異に関するしきい値も同様で、入力テキストに応じてその最適値が異なるのが普通である。これは、入力テキスト文書の性質によっては期待通りの動作をしないということであるから、実際応用上深刻な問題となる。

- [0012] 第2の従来技術では、話題のモデルを作成するために、事前に大規模なテキストコーパスを準備しなければならないという問題がある。しかもそのテキストコーパスは、話題ごとに分割済みであることが必須であり、しばしば話題のラベル(例えば「政治」、「スポーツ」、「経済」等)が付与されていることが要求される。このようなテキストコーパスを事前に準備するのには、当然時間と費用がかかる。しかも、第2の従来技術では、話題のモデルを作成するのに使用したテキストコーパスが、入力テキスト中の話題と同じ話題を含んでいること、すなわちドメイン(分野)が一致していることが必要となる。したがって、この従来技術の例の場合、入力テキストのドメインが未知の場合、またはドメインが頻繁に変化し得る場合、所望のテキスト分割結果を得ることは困難である。
- [0013] 本発明の目的は、従来よりも低コストかつ短時間にテキスト文書を話題ごとに分割できるようにすることにある。

また、他の目的は、テキスト文書のドメインに依存することなく、文書の性質によって、文書を話題ごとに分割できるようにすることにある。

課題を解決するための手段

- [0014] 上記目的を達成するために、本発明のテキスト処理方法は、テキスト文書を構成する各々の単語がどの話題に属するかを隠れ変数(Latent variable)に、各々の単語を観測変数(Observable variable)にそれぞれ対応付けた確率モデルを生成するステップと、生成された確率モデルを規定するモデルパラメータの初期値を出力するステップと、出力されたモデルパラメータの初期値と、処理対象のテキスト文書とともにづいて、このテキスト文書に応じたモデルパラメータを推定するステップと、推定されたモ

モデルパラメータにもとづいて、処理対象のテキスト文書を話題ごとに分割するステップとを備えることを特徴とする。

- [0015] また、本発明のテキスト処理装置は、テキスト文書を構成する各々の単語がどの話題に属するかを隠れ変数に、各々の単語を観測変数にそれぞれ対応付けた確率モデルを生成する仮モデル生成手段と、前記仮モデル生成手段によって生成された確率モデルを規定するモデルパラメータの初期値を出力するモデルパラメータ初期化手段と、前記モデルパラメータ初期化手段から出力されたモデルパラメータの初期値と、処理対象のテキスト文書とともに、このテキスト文書に応じたモデルパラメータを推定するモデルパラメータ推定手段と、前記モデルパラメータ推定手段によって推定されたモデルパラメータにもとづいて、処理対象のテキスト文書を話題ごとに分割するテキスト分割結果出力手段とを備えることを特徴とする。

発明の効果

- [0016] 本発明によれば、処理対象のテキスト文書の性質に応じてパラメータを調整する手間が少なく、事前に時間と費用をかけて大規模なテキストコーパスを準備する必要もなく、なおかつ処理対象のテキスト文書がどのような内容を含んでいるか、すなわちドメインに依存せずに、文書を精度よく話題ごとに分割することが可能となる。

図面の簡単な説明

- [0017] [図1]図1は、本発明の一実施例に係るテキスト処理装置の構成を示すブロック図である。

[図2]図2は、本発明の一実施例に係るテキスト処理装置の動作を説明するためのフローチャートである。

[図3]図3は、隠れマルコフモデルを説明するための概念図である。

[図4]図4は、本発明の他の実施例に係るテキスト処理装置の構成を示すブロック図である。

[図5]図5は、第1の従来技術を説明するための概念図である。

[図6]図6は、第2の従来技術を説明するための概念図である。

発明を実施するための最良の形態

- [0018] 第1の実施例

次に、本発明の第1の実施例について、図面を参照して詳細に説明する。

- [0019] 本実施例のテキスト処理装置は、図1に示すように、テキスト文書を入力するテキスト入力部101と、入力されたテキスト文書を格納するテキスト記憶部102と、テキスト文書の話題(意味的にまとまった部分)の推移を記述するモデルであって、テキスト文書の各々の単語がどの話題に属するかを隠れ変数(観測不可能な変数)に、テキスト文書の各々の単語を観測変数(観測可能な変数)にそれぞれ対応付けた、单一もしくは複数のモデルを生成する仮モデル生成部103と、仮モデル生成部103が生成した各モデルを規定する各モデルパラメータの値を初期化するモデルパラメータ初期化部104と、モデルパラメータ初期化部104によって初期化されたモデルとテキスト記憶部102に格納されたテキスト文書を使って、そのモデルのモデルパラメータを推定するモデルパラメータ推定部105と、モデルパラメータ推定部105が行ったパラメータ推定の結果を格納する推定結果記憶部106と、推定結果記憶部106に複数のモデルのパラメータ推定結果が格納されている場合にその中から1つのモデルのパラメータ推定結果を選択するモデル選択部107と、モデル選択部107が選択したモデルのパラメータ推定結果から入力テキスト文書の分割を行って結果を出力するテキスト分割結果出力部108を備える。各々の部は、それぞれ計算機上に記憶されたプログラムによって、またはこのプログラムが記録された記録媒体を読み取ることによって動作させることにより実現可能である。
- [0020] ここでテキスト文書とは、上述したように、例えば、磁気ディスク等の記録媒体に記録された任意の文字や単語などの並びである。あるいは、紙に印刷されたりタブレットに手書きされたりした文字列を光学的読取り装置(OCR)で読み取った結果や、人の発話で生じる音声波形信号を音声認識装置で認識した結果等も、テキスト文書である。さらに一般的には、毎日の天候の記録、店舗における商品の販売記録、コンピュータを操作した際のコマンドの記録、等々、時系列的に生成される記号の並びのほとんどは、テキスト文書の範疇に入る。
- [0021] 次に、本実施例のテキスト処理装置の動作を、図2を参照して詳細に説明する。
- [0022] テキスト入力部101から入力されたテキスト文書は、テキスト記憶部102に格納される(ステップ201)。ここでテキスト文書は、多数、例えばT個の単語が一列に並んだ单

語系列とし、以下では o_1, o_2, \dots, o_T と表すことにする。単語間にスペースのない日本語の場合は、テキスト文書に対して公知の形態素解析法を適用することにより、単語に分割すればよい。また、この単語列から、テキスト文書の話題とは直接関係のない助詞や助動詞などをあらかじめ取り除いて、名詞や動詞などの重要語のみの単語列としてもよい。これには、公知の形態素解析法によって各単語の品詞を求め、名詞、動詞、形容詞などを重要語として取り出すようにすればよい。さらには、入力テキスト文書が、音声信号を音声認識して得られた音声認識結果であり、かつ音声信号に一定時間以上継続する無音(発話休止)区間が存在する場合は、テキスト文書の対応する位置に<ポーズ>のような単語を含めてよい。同様に、入力テキスト文書が、紙文書をOCRにかけることによって得られた文字認識結果である場合には、<改行>のような単語をテキスト文書中の対応する位置に含めてよい。

- [0023] なお、通常の意味での単語系列(ユニグラム, unigram)の代わりに、隣接する単語の2つ組(バイグラム, bigram)、3つ組(トライグラム, trigram)、さらに一般的なn個組(n-gram)を一種の単語と考えて、その系列をテキスト記憶部102に格納してもよい。例えば2つ組での単語列の格納形式は $(o_1, o_2), (o_2, o_3), \dots, (o_{T-1}, o_T)$ となり、系列の長さはT-1である。
- [0024] 仮モデル生成部103は、入力されたテキスト文書を生成したと推測される单一もしくは複数の確率モデルを生成する。ここで確率モデルまたはモデルとは、一般にはグラフィカルモデルと呼ばれる、複数のノードとそれらを結ぶアーケとで表現されるモデル全般を指す。グラフィカルモデルには、マルコフモデルやニューラルネットワーク、ベイジアンネットなどが含まれる。本実施例においては、ノードがテキスト中に含まれる話題に対応する。また、モデルから生成されて観測される観測変数には、テキスト文書の構成要素であるところの単語が対応する。
- [0025] 本実施例では、モデルを隠れマルコフモデル(Hidden Markov ModelまたはHMM)とし、なおかつその構造は一方向型(left-to-right型)で、出力は上述の入力単語列に含まれる単語の系列(離散値)とする。Left-to-right型HMMでは、ノードの数を指定すればモデルの構造が一意に決定される。このモデルの概念図を図3に示す。HMMの場合特に、ノードのことを状態と呼ぶのが一般的である。図3の場合、ノード数、すな

わち状態数は4である。

- [0026] 仮モデル生成部103は、入力テキスト文書にいくつの話題が含まれているかに応じて、モデルの状態数を決定し、その状態数に応じてモデルすなわちHMMを生成する。例えば、入力テキスト文書に4個の話題が含まれているとわかっていれば、仮モデル生成部103は4状態のHMMを1つだけ生成する。また、入力テキスト文書に含まれる話題の数が未知の場合は、十分小さい状態数 N_{\min} のHMMから、十分大きい状態数 N_{\max} のHMMまでのすべての状態数のHMMを、各々1つずつ生成する(ステップ202、206、207)。ここでモデルを生成するとは、モデルを規定するパラメータの値を記憶するための記憶領域を記憶媒体上に確保する、という意味である。モデルを規定するパラメータについては後述する。
- [0027] 入力テキスト文書に含まれる各々の話題と入力テキスト文書の各々の単語との対応関係を隠れ変数とする。隠れ変数は単語毎に設定される。話題の数がNの場合には、隠れ変数は各々の単語がどの話題に属するかによって、1からNまでの値をとり得る。この隠れ変数がモデルの状態を表す。
- [0028] モデルパラメータ初期化部104は、仮モデル生成部103が生成したすべてのモデルについて、モデルを規定するパラメータの値を初期化する(ステップ203)。モデルを規定するパラメータは、上述のleft-to-right型離散HMMの場合、状態遷移確率 a_1, a_2, \dots, a_N 、および記号出力確率 $b_{1,j}, b_{2,j}, \dots, b_{N,j}$ とする。ここにNは状態数である。また $j=1, 2, \dots, L$ で、Lは入力テキスト文書に含まれる単語の種類数、すなわち語彙数である。
- 状態遷移確率 a_i は、状態iから状態i+1に遷移する確率であり、 $0 < a_i \leq 1$ でなければならない。よって、状態iから再度状態iに戻る確率は $1-a_i$ となる。また、記号出力確率 $b_{i,j}$ は、ある一度の状態遷移の後に、状態iに至ったとして、インデックスjで指定される単語が出力される確率である。すべての状態 $i=1, 2, \dots, N$ において、記号出力確率の総和 $b_{i,1} + b_{i,2} + \dots + b_{i,L}$ は1でなければならない。
- [0029] モデルパラメータ初期化部104は、状態数Nのモデルに対して、例えば上述の各パラメータの値を $a_i=N/T$ 、 $b_{i,j}=1/L$ のように設定する。この初期値の与え方に決まったやり方ではなく、上述の確率の条件さえ満たしていれば、いろいろな方法があり得る。ここ

で述べた方法はほんの一例である。

[0030] モデルパラメータ推定部105は、モデルパラメータ初期化部104によって初期化された单一もしくは複数のモデルを順次受け取り、モデルが入力テキスト文書 o_1, o_2, \dots, o_T を生成する確率、すなわち尤度がなるべく高くなるように、モデルパラメータを推定する(ステップ204)。これには公知の最尤推定法、特に、反復計算を基本とする期待値最大化法(EM(expectation-maximization)法)を用いることができる。すなわち、例えば「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第129～134頁」(文献4)に記載されているように、その時点で得られているパラメータ値 a_i, b_{ij} を用いて、式(2)のような漸化式によって前向き変数 α_t (i)および後向き変数 β_t (i)を $t=1, 2, \dots, T, i=1, 2, \dots, N$ にわたって計算し、さらに式(3)に従ってパラメータ値を再計算する。再計算されたパラメータ値を用いて再度式(2)および式(3)を計算する。以下、収束するまで十分な回数これをくり返す。ただしここに δ_{ij} はクロネッカーデルタ、すなわち、 $i=j$ なら1、そうでなければ0をとる。

[0031] [数2]

$$\begin{aligned}\alpha_1(i) &= b_{1,o_1} \delta_{1,i}, \quad \alpha_t(i) = \alpha_{t-1} b_{i,o_t} \alpha_{t-1}(i-1) + (1 - \alpha_{t-1}) b_{i,o_t} \alpha_{t-1}(i), \\ \beta_T(i) &= \alpha_N \delta_{N,i}, \quad \beta_t(i) = (1 - \alpha_t) b_{i,o_{t+1}} \beta_{t+1}(i) + \alpha_t b_{i+1,o_{t+1}} \beta_{t+1}(i+1).\end{aligned}\dots(2)$$

[0032] [数3]

$$\begin{aligned}a_i &\leftarrow \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_i b_{i+1,o_t} \beta_{t+1}(i+1)}{\sum_{t=1}^{T-1} \alpha_t(i) (1 - a_t) b_{i,o_t} \beta_{t+1}(i) + \sum_{t=1}^{T-1} \alpha_t(i) a_i b_{i+1,o_t} \beta_{t+1}(i+1)}, \\ b_{ij} &\leftarrow \frac{\sum_{t=1}^T \alpha_t(i) \beta_t(i) \delta_{j,o_t}}{\sum_{t=1}^T \alpha_t(i) \beta_t(i)}.\end{aligned}\dots(3)$$

[0033] モデルパラメータ推定部105におけるパラメータ推定の反復計算の収束判定を行うには、尤度の上昇量をみればよい。すなわち、上述の反復計算によって尤度の上昇がみられなくなれば、その時点で反復計算を終了すればよい。ここで、尤度は $\alpha_1(1)$ $\beta_1(1)$ として得られる。モデルパラメータ推定部105は、反復計算を終了した時点で、

モデルパラメータ a_{ij} 、 b_{ij} と、前向きおよび後向き変数 $\alpha_t(i)$ 、 $\beta_t(i)$ を、モデル(HMM)の状態数と対にして、推定結果記憶部106に格納する(ステップ205)。

- [0034] モデル選択部107は、モデルパラメータ推定部105で状態数ごとに得られたパラメータ推定結果を推定結果記憶部106から受け取り、各モデルの確からしさを計算し、もっとも確からしいモデルを1つ選択する(ステップ208)。モデルの確からしさは、公知の赤池情報量基準(AIC(Akaike's Information Criterion))や最小記述長基準(MDL(Minimum Description Length)基準)などに基づいて計算することができる。赤池情報量基準、最小記述長基準については、例えば「1994年12月、岩波書店、岩波講座応用数学[対象11]、韓太舜他著、情報と符号化の数理、第249～275頁」(文献5)に記載がある。例えばAICによれば、パラメータ推定収束後の対数尤度 $\log(\alpha_1(1) \beta_1(1))$ とモデルパラメータ数NLの差が最大となるモデルが選択される。また、MDLによれば、近似的に、対数尤度を符号反転した $-\log(\alpha_1(1) \beta_1(1))$ と、モデルパラメータ数と入力テキスト文書の単語系列長の平方根との積NL × log(T)/2の和が最小となるモデルが選択される。なお、AICでもMDLでも、モデルパラメータ数NLに関わる項に、経験的に決まる定数係数をかけて、選択されるモデルを意図的に調整する操作が一般的に行われているが、本実施例でもそのような操作は行って差し支えない。

- [0035] テキスト分割結果出力部108は、モデル選択部107によって選択された状態数Nのモデルに対応するモデルパラメータ推定結果を推定結果記憶部106から受け取り、この推定結果における入力テキスト文書に対する話題ごとの分割結果を算出する(ステップ209)。

状態数Nのモデルによる分割は、入力テキスト文書 o_1, o_2, \dots, o_T をN個の区間に分割する。分割結果は、まず式(4)に従って、確率的に計算される。式(4)は、入力テキスト文書中の単語 o_t が第i番目の話題区間に割り当てられる確率を示す。最終的な分割結果は、 $P(z_t=i | o_1, o_2, \dots, o_T)$ が最大となるiを $t=1, 2, \dots, T$ にわたって求めることで得られる。

- [0036] [数4]

$$P(z_t = i | o_1, o_2, \dots, o_T) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)} \quad \dots (4)$$

[0037] なお、ここではモデルパラメータ推定部105は、最尤推定法を用いて、すなわち式(3)を用いて、パラメータを逐次更新したが、最尤推定法の他に、最大事後確率推定(MAP(Maximum A Posteriori)推定)を用いることもできる。最大事後確率推定については、例えば「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第166～169頁」(文献6)に記載がある。最大事後確率推定の場合、例えばモデルパラメータの事前分布に共役事前分布を用いると、 a_i の事前分布はベータ分布 $\log p(a_i | \kappa_0, \kappa_1) = (\kappa_0 - 1) \times \log(1 - a_i) + (\kappa_1 - 1) \times \log(a_i) + \text{const}$ 、 b_{ij} の分布はディレクレ分布 $\log p(b_{i,1}, b_{i,2}, \dots, b_{i,L} | \lambda_1, \lambda_2, \dots, \lambda_L) = (\lambda_1 - 1) \times \log(b_{i,1}) + (\lambda_2 - 1) \times \log(b_{i,2}) + \dots + (\lambda_L - 1) \times \log(b_{i,L}) + \text{const}$ と表される。ただし $\kappa_0, \kappa_1, \lambda_1, \lambda_2, \dots, \lambda_L$ およびconstは定数である。このとき、最尤推定の式(3)に相当する最大事後確率推定のパラメータ更新式は、式(5)のように表される。

[0038] [数5]

$$a_i \leftarrow \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_i b_{i+1,o_t} \beta_{t+1}(i+1) + \kappa_1 - 1}{\sum_{t=1}^{T-1} \alpha_t(i)(1 - a_i) b_{i,o_t} \beta_{t+1}(i) + \kappa_0 - 1 + \sum_{t=1}^{T-1} \alpha_t(i) a_i b_{i+1,o_t} \beta_{t+1}(i+1) + \kappa_1 - 1}, \quad \dots (5)$$

$$b_{ij} \leftarrow \frac{\sum_{t=1}^T \alpha_t(i) \beta_t(i) \delta_{j,o_t} + \lambda_j - 1}{\sum_{t=1}^T \alpha_t(i) \beta_t(i) + \sum_{k=1}^L (\lambda_k - 1)}.$$

[0039] なお、ここまで述べた本実施例においては、記号出力確率 b_{ij} が状態と対応付けられている。すなわち、単語がHMMの各状態(ノード)から発生するとするモデルを用いている。しかし、単語が状態遷移(アーク)から発生するとするモデルを用いることも可能である。例えば入力テキストが紙文書のOCR結果であったり、音声信号の音声認識結果であったりする場合、単語が状態遷移から発生するようなモデルは便利である。なぜなら、音声信号における発話休止や、紙文書における改行などを意味する単

語、すなわち<ポーズ>や<改行>などが含まれたテキスト文書の場合は、状態iからi+1への状態遷移から発生する単語が必ず<ポーズ>や<改行>であるように、記号出力確率を固定しておけば、本実施例によって入力テキスト文書から検出される話題境界には、必ず<ポーズ>や<改行>が当てはまるようになります。また、仮に入力テキスト文書がOCR結果や音声認識結果ではなくとも、単語が状態遷移から発生するモデルで、状態iからi+1への状態遷移から、「では」、「次に」、「さて」などといった、話題の切り替わりと関連の深い単語が発生するように記号出力確率を設定しておけば、検出される話題境界には「では」、「次に」、「さて」などの単語が現れやすくなる。

[0040] 第2の実施例

次に、本発明の第2の実施例について、図面を参照して詳細に説明する。

[0041] 本実施例は、第1の実施例と同じく、図1のブロック図で示される。すなわち、本実施例は、テキスト文書を入力するテキスト入力部101と、入力されたテキスト文書を格納するテキスト記憶部102と、テキスト文書の話題の推移を記述するモデルであって、テキスト文書の各々の単語がどの話題に属するかを隠れ変数に、テキスト文書の各々の単語を観測変数にそれぞれ対応付けた、单一もしくは複数のモデルを生成する仮モデル生成部103と、仮モデル生成部103が生成した各モデルを規定する各モデルパラメータの値を初期化するモデルパラメータ初期化部104と、モデルパラメータ初期化部104によって初期化されたモデルとテキスト記憶部102に格納されたテキスト文書を使ってモデルパラメータを推定するモデルパラメータ推定部105と、モデルパラメータ推定部105が行ったパラメータ推定の結果を格納する推定結果記憶部106と、推定結果記憶部106に複数のモデルのパラメータ推定結果が格納されている場合にその中から1つのモデルのパラメータ推定結果を選択するモデル選択部107と、モデル選択部107が選択したモデルのパラメータ推定結果から入力テキスト文書の分割を行つて結果を出力するテキスト分割結果出力部108を備える。各々の部は、それぞれ計算機上に記憶されたプログラムによって、またはこのプログラムが記録された記録媒体を読み取ることによって動作させることにより実現可能である。

[0042] 次に、本実施例の動作について、順を追つて説明する。

[0043] テキスト入力部101、テキスト記憶部102および仮モデル生成部103は、それぞれ先

に述べた第1の実施例におけるテキスト入力部101、テキスト記憶部102および仮モデル生成部103と同一の動作をする。テキスト記憶部102が入力テキスト文書を、単語の列、あるいは隣接する単語の2つ組、3つ組、もしくは一般のn個組の列として格納することや、入力テキスト文書に単語間スペースのない日本語の場合、公知の形態素解析法を適用することで、単語列として扱うことができることなども、第1の実施例と同様である。

- [0044] モデルパラメータ初期化部104は、仮モデル生成部103が生成したすべてのモデルについて、モデルを規定するパラメータの値を初期化する。モデルは、第1の実施例と同様、left-to-right型離散HMMであるが、さらにタイドミクスチャ(tied-mixture)HMMであるとする。すなわち、状態iからの記号出力が、M個の記号出力確率 $b_{1,j}, b_{2,j}, \dots, b_{M,j}$ の線形結合 $c_{i,1}b_{1,j} + c_{i,2}b_{2,j} + \dots + c_{i,M}b_{M,j}$ であり、 $b_{i,j}$ の値は全状態にわたって共通とする。Mは一般には状態数Nよりも小さい、任意の自然数である。タイドミクスチャHMMについては、例えば「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第280～281頁」(文献7)に記載がある。タイドミクスチャ(tied-mixture)HMMのモデルパラメータは、状態遷移確率 a_i 、全状態で共通の記号出力確率 $b_{j,k}$ 、および記号出力確率に対する重み係数 $c_{i,j}$ である。ここで、 $i=1, 2, \dots, N$ で、Nは状態数である。 $j=1, 2, \dots, M$ で、Mは話題の種類数。また $k=1, 2, \dots, L$ で、Lは入力テキスト文書に含まれる単語の種類数、すなわち語彙数である。状態遷移確率 a_i は、第1の実施例と同様、状態iから状態i+1に遷移する確率である。記号出力確率 $b_{j,k}$ は、話題jにおいて、インデックスkで指定される単語が出力される確率である。また重み係数 $c_{i,j}$ は、状態iにおいて話題jが発生する確率である。第1の実施例と同様、記号出力確率の総和 $b_{j,1} + b_{j,2} + \dots + b_{j,L}$ は1でなければならない。また、重み係数の総和 $c_{i,1} + c_{i,2} + \dots + c_{i,L}$ も1でなければならない。
- [0045] モデルパラメータ初期化部104は、状態数Nのモデルに対して、例えば上述の各パラメータの値を $a_i = N/T$ 、 $b_{j,k} = 1/L$ 、 $c_{i,j} = 1/M$ のように設定する。この初期値の与え方に決まったやり方ではなく、上述の確率の条件さえ満たしていれば、いろいろな方法があり得る。ここで述べた方法はほんの一例である。
- [0046] モデルパラメータ推定部105は、モデルパラメータ初期化部104によって初期化され

た单一もしくは複数のモデルを順次受け取り、モデルが入力テキスト文書 o_1, o_2, \dots, o_T を生成する確率、すなわち尤度がなるべく高くなるように、モデルパラメータを推定する。これには、第1の実施例と同様、期待値最大化法(EM法)を用いることができる。すなわち、その時点で得られているパラメータ値 $a_i, b_{j,k}, c_{i,j}$ を用いて、式(6)のような漸化式によって前向き変数 $\alpha_t(i)$ および後向き変数 $\beta_t(i)$ を $t=1, 2, \dots, T, i=1, 2, \dots, N$ にわたって計算し、さらに式(7)に従ってパラメータ値を再計算する。再計算されたパラメータ値を用いて再度式(6)および式(7)を計算する。以下、収束するまで十分な回数これをくり返す。ただしここに δ_{ij} はクロネッカーデルタ、すなわち、 $i=j$ なら1、そうでなければ0をとる。

[0047] [数6]

$$\begin{aligned} \alpha_1(i) &= \sum_{j=1}^M c_{1,j} b_{j,o_1} \delta_{1,i}, \quad \alpha_t(i) = \sum_{j=1}^M \left\{ a_{t-1} c_{i,j} b_{j,o_t} \alpha_{t-1}(i-1) + (1-a_t) c_{i,j} b_{j,o_t} \alpha_{t-1}(i) \right\}, \\ \beta_T(i) &= a_N \delta_{N,i}, \quad \beta_t(i) = \sum_{j=1}^M \left\{ (1-a_t) c_{i,j} b_{j,o_{t+1}} \beta_{t+1}(i) + a_t c_{i+1,j} b_{j,o_{t+1}} \beta_{t+1}(i+1) \right\} \end{aligned} \quad \cdots (6)$$

[0048] [数7]

$$\begin{aligned} a_i &\leftarrow \frac{\sum_{t=1}^{T-1} \sum_{j=1}^M \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1)}{\sum_{t=1}^{T-1} \sum_{j=1}^M \left\{ \alpha_t(i)(1-a_t) c_{i,j} b_{j,o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1) \right\}}, \\ b_{ij} &\leftarrow \frac{\sum_{t=1}^T \sum_{i=1}^N \left\{ \alpha_t(i)(1-a_t) c_{i,j} b_{j,o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1) \right\}}{\sum_{t=1}^T \sum_{i=1}^N \sum_{k=1}^L \left\{ \alpha_t(i)(1-a_t) c_{i,j} b_{j,k} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j} b_{j,k} \beta_{t+1}(i+1) \right\}}, \\ c_{ij} &\leftarrow \frac{\sum_{t=1}^T \left\{ \alpha_t(i)(1-a_t) c_{i,j} b_{j,o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1) \right\}}{\sum_{j'=1}^M \sum_{t=1}^T \left\{ \alpha_t(i)(1-a_t) c_{i,j'} b_{j',o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j'} b_{j',o_t} \beta_{t+1}(i+1) \right\}}. \end{aligned} \quad \cdots (7)$$

[0049] モデルパラメータ推定部105におけるパラメータ推定の反復計算の収束判定を行うには、尤度の上昇量をみればよい。すなわち、上述の反復計算によって尤度の上昇がみられなくなれば、その時点で反復計算を終了すればよい。ここに、尤度は $\alpha_1(1) \beta_1(1)$ として得られる。モデルパラメータ推定部105は、反復計算を終了した時点で、

モデルパラメータ $a_{i,j,k}$ 、 $b_{j,k}$ 、 $c_{i,j}$ と、前向きおよび後向き変数 $\alpha_t(i)$ 、 $\beta_t(i)$ を、モデル(HMM)の状態数と対にして、推定結果記憶部106に格納する。

[0050] モデル選択部107は、第1の実施例と同様、モデルパラメータ推定部105で状態数ごとに得られたパラメータ推定結果を推定結果記憶部106から受け取り、各モデルの確からしさを計算し、もっとも確からしいモデルを1つ選択する。モデルの確からしさは、公知の赤池情報量基準(AIC)や最小記述長基準(MDL基準)などに基づいて計算することができる。

また、第1の実施例と同様、AICでもMDLでも、モデルパラメータ数NLに関わる項に、経験的に決まる定数係数をかけて、選択されるモデルを意図的に調整する操作も行って差し支えない。

[0051] テキスト分割結果出力部108は、第1の実施例におけるテキスト分割結果出力部108と同様、モデル選択部107によって選択された状態数すなわち話題数Nのモデルに対応するモデルパラメータ推定結果を推定結果記憶部106から受け取り、この推定結果における入力テキスト文書に対する話題ごとの分割結果を算出する。最終的な分割結果は、式(4)に従って、 $P(z_t=i | o_1, o_2, \dots, o_T)$ が最大となる i を $t=1, 2, \dots, T$ にわたって求めることで得られる。

[0052] なお、モデルパラメータ推定部105は、第1の実施例と同様、最尤推定法の代わりに最大事後確率推定(MAP推定)法によってモデルパラメータを推定してもよい。

[0053] 第3の実施例

次に、本発明の第3の実施例について、図面を参照して説明する。

[0054] 本実施例は、第1および第2の実施例の例と同じく、図1のブロック図で示される。すなわち、本実施例は、テキスト文書を入力するテキスト入力部101と、入力されたテキスト文書を格納するテキスト記憶部102と、テキスト文書の話題の推移を記述するモデルであって、テキスト文書の各々の単語がどの話題に属するかを隠れ変数に、テキスト文書の各々の単語を観測変数にそれぞれ対応付けた、单一もしくは複数のモデルを生成する仮モデル生成部103と、仮モデル生成部103が生成した各モデルを規定する各モデルパラメータの値を初期化するモデルパラメータ初期化部104と、モデルパラメータ初期化部104によって初期化されたモデルとテキスト記憶部102に格納され

たテキスト文書を使ってモデルパラメータを推定するモデルパラメータ推定部105と、モデルパラメータ推定部105が行ったパラメータ推定の結果を格納する推定結果記憶部106と、推定結果記憶部106に複数のモデルのパラメータ推定結果が格納されている場合にその中から1つのモデルのパラメータ推定結果を選択するモデル選択部107と、モデル選択部107が選択したモデルのパラメータ推定結果から入力テキスト文書の分割を行って結果を出力するテキスト分割結果出力部108を備える。各々の部は、それぞれ計算機上に記憶されたプログラムによって、またはこのプログラムが記録された記録媒体を読み取ることによって動作させることにより実現可能である。

- [0055] 次に、本実施例の動作について、順を追って説明する。
- [0056] テキスト入力部101、テキスト記憶部102および仮モデル生成部103は、それぞれ先に述べた第1および第2の実施例におけるテキスト入力部101、テキスト記憶部102および仮モデル生成部103と同一の動作をする。テキスト記憶部102が入力テキスト文書を、単語の列、あるいは隣接する単語の2つ組、3つ組、もしくは一般のn個組の列として格納することや、入力テキスト文書に単語間スペースのない日本語の場合、公知の形態素解析法を適用することで、単語列として扱うことができるなども、本発明の第1および第2の実施例と同様である。
- [0057] モデルパラメータ初期化部104は、仮モデル生成部103が生成した单一または複数のモデル各々について、モデルパラメータ、すなわち状態遷移確率 a_i および記号出力確率 b_{ij} を確率変数として、ある種の分布を仮定し、それらの分布を規定するパラメータの値を初期化する。以下では、モデルパラメータの分布を規定するパラメータを、元のパラメータに対してメタパラメータと呼ぶことにする。つまり、モデルパラメータ初期化部104はメタパラメータの初期化を行う。本実施例では、状態遷移確率 a_i および記号出力確率 b_{ij} の分布として、それぞれベータ分布 $\log p(a_i | \kappa_{0,i}, \kappa_{1,i}) = (\kappa_{0,i} - 1) \times \log(1 - a_i) + (\kappa_{1,i} - 1) \times \log(a_i) + \text{const}$ 、ディレクレ分布 $\log p(b_{i,1}, b_{i,2}, \dots, b_{i,L} | \lambda_{i,1}, \lambda_{i,2}, \dots, \lambda_{i,L}) = (\lambda_{i,1} - 1) \times \log(b_{i,1}) + (\lambda_{i,2} - 1) \times \log(b_{i,2}) + \dots + (\lambda_{i,L} - 1) \times \log(b_{i,L}) + \text{const}$ を使用する。メタパラメータは $\kappa_{0,i}, \kappa_{1,i}, \lambda_{ij}$ である。ここで、 $i=1, 2, \dots, N, j=1, 2, \dots, L$ である。モデルパラメータ初期化部104は、例えば $\kappa_{0,i} = \kappa_0, \kappa_{1,i} = \kappa_1, \lambda_{ij} = \lambda_0$ 、ただし $\kappa_0 = \varepsilon(1-N/T) + 1, \kappa_1 = \varepsilon N/T + 1, \lambda_0 = \varepsilon/L + 1$ 、というようにメタパラメータを初期

化する。 ϵ としては、0.01などのように適当な正数を当てる。なお、初期値の与え方に決まったやり方はなく、いろいろな方法があり得る。

この初期化方法はほんの一例である。

[0058] モデルパラメータ推定部105は、モデルパラメータ初期化部104によって初期化された单一もしくは複数のモデルを順次受け取り、モデルが入力テキスト文書 o_1, o_2, \dots, o_T を生成する確率、すなわち尤度がなるべく高くなるように、メタパラメータを推定する。これにはベイズ推定法から導出される公知の変分ベイズ法を用いることができる。すなわち、例えば「2002年7月、電子情報通信学会誌、第85巻、第7号、第504～509頁、上田、ベイズ学習[III]－変分ベイズ学習の基礎－」(文献8)に記載があるように、その時点で得られているメタパラメータ値 $\kappa_{0,i}, \kappa_{1,i}, \lambda_{ij}$ を用いて、式(8)のような漸化式によって前向き変数 $\alpha_t(i)$ および後向き変数 $\beta_t(i)$ を $t=1, 2, \dots, T, i=1, 2, \dots, N$ にわたって計算し、さらに式(9)に従ってメタパラメータ値を再計算する。再計算されたパラメータ値を用いて、再度式(8)および式(9)を計算する。以下、収束するまで十分な回数これをくり返す。ただしここに、 δ_{ij} はクロネッカーデルタ、すなわち、 $i=j$ なら1、そうでなければ0をとる。また、 $\Psi(x)=d(\log \Gamma(x))/dx$ で、 $\Gamma(x)$ はガンマ関数である。

[0059] [数8]

$$\begin{aligned}
 \alpha_1(i) &= \exp(B_{i,o_1})\delta_{1,i}, \\
 \alpha_t(i) &= \alpha_{t-1}(i-1)\exp(A_{1,t-1} + B_{i,o_t}) + \alpha_{t-1}(i)\exp(A_{0,i} + B_{i,o_t}), \\
 \beta_T(i) &= \exp(A_{1,N})\delta_{N,i}, \\
 \beta_t(i) &= \beta_{t+1}(i)\exp(A_{0,i} + B_{i,o_{t+1}}) + \beta_{t+1}(i+1)\exp(A_{1,i} + B_{i+1,o_{t+1}}) \\
 \text{ただし} \\
 A_{0,i} &= \Psi(\kappa_{0,i}) - \Psi(\kappa_{0,i} + \kappa_{1,i}), \\
 A_{1,i} &= \Psi(\kappa_{1,i}) - \Psi(\kappa_{0,i} + \kappa_{1,i}), \\
 B_{ik} &= \Psi(\lambda_{ik}) - \Psi\left(\sum_{j=1}^L \lambda_{ij}\right).
 \end{aligned} \quad \cdots (8)$$

[0060] [数9]

$$\kappa_{0,i} \leftarrow \kappa_0 + \sum_{t=1}^{T-1} \overline{z_{t,i} z_{t+1,i}}, \quad \kappa_{1,i} \leftarrow \kappa_1 + \sum_{t=1}^{T-1} \overline{z_{t,i} z_{t+1,i+1}} + \delta_{N,i}, \quad \lambda_{ik} \leftarrow \lambda_0 + \sum_{t=1}^{T-1} \overline{z_{t,i}} \delta_{k,o_t}. \quad \dots (9)$$

ただし

$$\begin{aligned} \overline{z_{t,i}} &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)}, \\ \overline{z_{t,i} z_{t+1,i}} &= \frac{\alpha_t(i)\exp(A_{0,i} + B_{i,o_{t+1}})\beta_{t+1}(i)}{\sum_{j=1}^N \sum_{s=\{0,1\}} \alpha_t(j)\exp(A_{s,j} + B_{j+s,o_{t+1}})\beta_{t+1}(j+s)}, \\ \overline{z_{t,i} z_{t+1,i+1}} &= \frac{\alpha_t(i)\exp(A_{1,i} + B_{i+1,o_{t+1}})\beta_{t+1}(i+1)}{\sum_{j=1}^N \sum_{s=\{0,1\}} \alpha_t(j)\exp(A_{s,j} + B_{j+s,o_{t+1}})\beta_{t+1}(j+s)}. \end{aligned}$$

[0061] モデルパラメータ推定部105におけるパラメータ推定の反復計算の収束判定は、近似的尤度の上昇量をみればよい。すなわち、上述の反復計算によって近似的尤度の上昇がみられなくなれば、その時点で反復計算を終了すればよい。ここで、近似的尤度とは、前向き変数と後向き変数の積 $\alpha_1(1) \beta_1(1)$ として得られる。モデルパラメータ推定部105は、反復計算を終了した時点で、メタパラメータ $\kappa_{0,i}$, $\kappa_{1,i}$, $\lambda_{i,j}$ と、前向きおよび後向き変数 $\alpha_t(i)$, $\beta_t(i)$ を、モデル(HMM)の状態数Nと対にして、推定結果記憶部106に格納する。

[0062] なお、モデルパラメータ推定部105におけるメタパラメータのベイズ推定法としては、上述の変分ベイズ法以外にも、公知のマルコフ連鎖モンテカルロ法やラプラス近似法など、任意の方法を使うことができる。本実施例は、変分ベイズ法に限定されるものではない。

[0063] モデル選択部107は、モデルパラメータ推定部105で状態数ごとに得られたパラメータ推定結果を推定結果記憶部106から受け取り、各モデルの確からしさを計算し、もつとも確からしいモデルを1つ選択する。モデルの確からしさは、例えば上述した変分ベイズ法の枠組みでは、公知のベイズ的基準(ベイズ事後確率)を使用することができる。ベイズ的基準は式(10)で計算可能である。式(10)において $P(N)$ は状態数すなわち話題数Nの事前確率で、あらかじめ何らかの方法で定めておく。取り立てて理由がなければ、 $P(N)$ は一定値でよい。逆に、特定の状態数が起こりやすい、あるいは起こりにくいということが事前にわかっている場合は、特定の状態数に対応する

$P(N)$ を大きく、あるいは小さく設定する。また、式(10)に現れるメタパラメータ $\kappa_{0,i}$, $\kappa_{1,i}$, $\lambda_{i,j}$ と、前向きおよび後向き変数 $\alpha_t(i)$, $\beta_t(i)$ としては、状態数Nに対応するものを推定結果記憶部106から取得して用いる。

[0064] [数10]

$$\begin{aligned}
 & P(N)\alpha_1(1)\beta_1(1) \\
 & \times \exp \left\{ \sum_{i=1}^N (\kappa_{0,i} - \kappa_0) (\Psi(\kappa_{0,i} + \kappa_{1,i}) - \Psi(\kappa_{0,i})) + \sum_{i=1}^N (\kappa_{1,i} - \kappa_1) (\Psi(\kappa_{0,i} + \kappa_{1,i}) - \Psi(\kappa_{1,i})) \right\} \\
 & \times \exp \left\{ \sum_{i=1}^N \sum_{k=1}^L (\lambda_{ij} - \lambda_0) \left(\Psi \left(\sum_{j=1}^L \lambda_{ij} \right) - \Psi(\lambda_{ik}) \right) \right\} \\
 & \times \prod_{i=1}^N \left\{ \frac{\Gamma(\kappa_0 + \kappa_1)}{\Gamma(\kappa_{0,i} + \kappa_{1,i})} \frac{\Gamma(\kappa_{0,i})\Gamma(\kappa_{1,i})}{\Gamma(\kappa_0)\Gamma(\kappa_1)} \frac{\Gamma \left(\sum_{j=1}^L \lambda_{ij} \right)}{\Gamma \left(\sum_{j=1}^L \lambda_{ij} \right)} \prod_{j=1}^L \frac{\Gamma(\lambda_{ij})}{\Gamma(\lambda_0)} \right\} \quad \cdots (10)
 \end{aligned}$$

[0065] テキスト分割結果出力部108は、上述の第1および第2の実施例におけるテキスト分割結果出力部108と同様、モデル選択部107によって選択された状態数すなわち話題数Nのモデルに対応するモデルパラメータ推定結果を推定結果記憶部106から受け取り、この推定結果における入力テキスト文書に対する話題ごとの分割結果を算出する。最終的な分割結果は、式(4)に従って、 $P(z_t=i | o_1, o_2, \dots, o_T)$ が最大となる i を $t=1, 2, \dots, T$ にわたって求めることで得られる。

[0066] なお、本実施例でも、上述した第2の実施例と同様、通常のleft-to-right型HMMの代わりに、タイドミクスチャ(tied-mixture)型のleft-to-right型HMMを生成、初期化、パラメータ推定するように、仮モデル生成部103、モデルパラメータ初期化部104、モデルパラメータ推定部105をそれぞれ構成することが可能である。

[0067] 第4の実施例

次に、本発明の第4の実施例について、図面を参照して詳細に説明する。

[0068] 図4を参照すると、本発明の第4の実施例は、テキスト処理プログラム605を記録した記録媒体601を備える。この記録媒体601はCD-ROM、磁気ディスク、半導体メモリその他の記録媒体であってよく、ネットワークを介して流通する場合も含む。テキスト処

理プログラム605は記録媒体601からデータ処理装置(コンピュータ)602に読み込まれ、データ処理装置602の動作を制御する。

[0069] 本実施例としては、データ処理装置602はテキスト処理プログラム605の制御により、第1、第2、もしくは第3の実施例におけるテキスト入力部101、仮モデル生成部103、モデルパラメータ初期化部104、モデルパラメータ推定部105、モデル選択部107、テキスト分割結果出力部108による処理と同一の処理を実行して、第1、第2、もしくは第3の実施例におけるテキスト記憶部102、推定結果記憶部106とそれぞれ同等の情報を有するテキスト記録媒体603、モデルパラメータ推定結果記録媒体604を参照することによって、入力されたテキスト文書に対する話題ごとの分割結果を出力する。

請求の範囲

- [1] テキスト文書を構成する各々の単語がどの話題に属するかを隠れ変数に、各々の単語を観測変数にそれぞれ対応付けた確率モデルを生成するステップと、
生成された確率モデルを規定するモデルパラメータの初期値を出力するステップと
、
出力されたモデルパラメータの初期値と、処理対象のテキスト文書とともにとづいて、
このテキスト文書に応じたモデルパラメータを推定するステップと、
推定されたモデルパラメータにもとづいて、処理対象のテキスト文書を話題ごとに分
割するステップと
を備えることを特徴とするテキスト処理方法。
- [2] 請求項1に記載のテキスト処理方法において、
確率モデルを生成する前記ステップは、複数の確率モデルを生成するステップを
備え、
モデルパラメータの初期値を出力する前記ステップは、複数の確率モデルのそれ
ぞれのモデルパラメータの初期値を出力するステップを備え、
モデルパラメータを推定する前記ステップは、複数の確率モデルのそれぞれのモ
デルパラメータを推定するステップを備え、
さらに、推定された複数のモデルパラメータにもとづいて、複数の確率モデルの中
から、テキスト文書を分割する前記ステップで処理を行う確率モデルを選択するステ
ップを備えることを特徴とするテキスト処理方法。
- [3] 請求項1に記載のテキスト処理方法において、
確率モデルは、隠れマルコフモデルであることを特徴とするテキスト処理方法。
- [4] 請求項3に記載のテキスト処理方法において、
隠れマルコフモデルは、一方向型の構造を有することを特徴とするテキスト処理方
法。
- [5] 請求項3に記載のテキスト処理方法において、
隠れマルコフモデルは、離散出力型であることを特徴とするテキスト処理方法。
- [6] 請求項1に記載のテキスト処理方法において、

モデルパラメータを推定する前記ステップは、最尤推定および最大事後確率推定のいずれかを用いてモデルパラメータを推定するステップを備えることを特徴とするテキスト処理方法。

[7] 請求項1に記載のテキスト処理方法において、

モデルパラメータの初期値を出力する前記ステップは、モデルパラメータを確率変数とする分布を仮定し、この分布を規定するメタパラメータの初期値を出力するステップを備え、

モデルパラメータを推定する前記ステップは、出力されたメタパラメータの初期値と、処理対象のテキスト文書とともにとづいて、このテキスト文書に応じたメタパラメータを推定するステップを備えることを特徴とするテキスト処理方法。

[8] 請求項7に記載のテキスト処理方法において、

メタパラメータを推定する前記ステップは、ベイズ推定を用いてメタパラメータを推定するステップを備えることを特徴とするテキスト処理方法。

[9] 請求項2に記載のテキスト処理方法において、

確率モデルを選択する前記ステップは、赤池情報量基準、最小記述長基準およびベイズ事後確率のいずれかを用いて確率モデルを選択するステップを備えることを特徴とするテキスト処理方法。

[10] テキスト文書を構成する各々の単語がどの話題に属するかを隠れ変数に、各々の単語を観測変数にそれぞれ対応付けた確率モデルを生成するステップと、

生成された確率モデルを規定するモデルパラメータの初期値を出力するステップと、

、

出力されたモデルパラメータの初期値と、処理対象のテキスト文書とともにとづいて、このテキスト文書に応じたモデルパラメータを推定するステップと、

推定されたモデルパラメータにもとづいて、処理対象のテキスト文書を話題ごとに分割するステップと

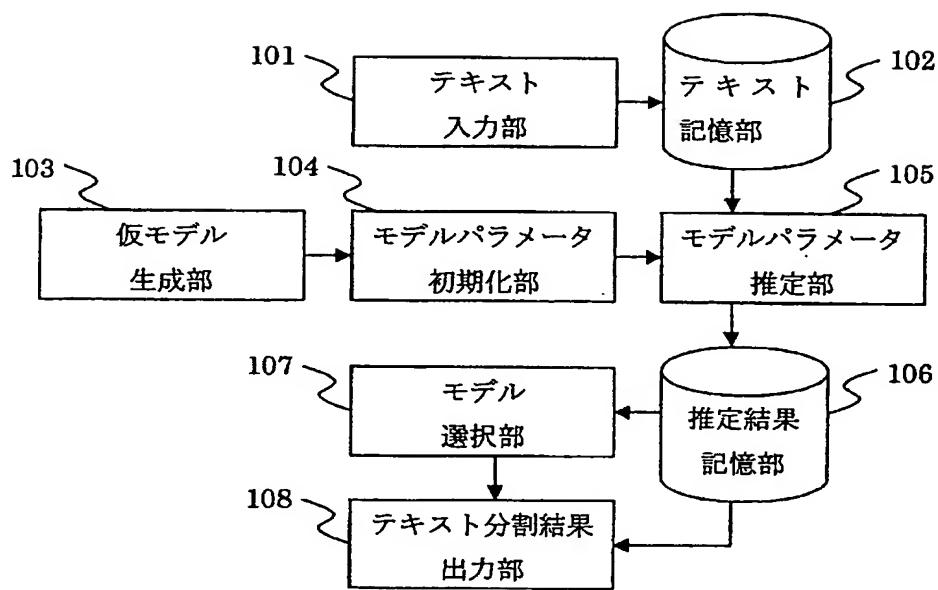
をコンピュータに実行させるためのプログラム。

[11] テキスト文書を構成する各々の単語がどの話題に属するかを隠れ変数に、各々の単語を観測変数にそれぞれ対応付けた確率モデルを生成するステップと、

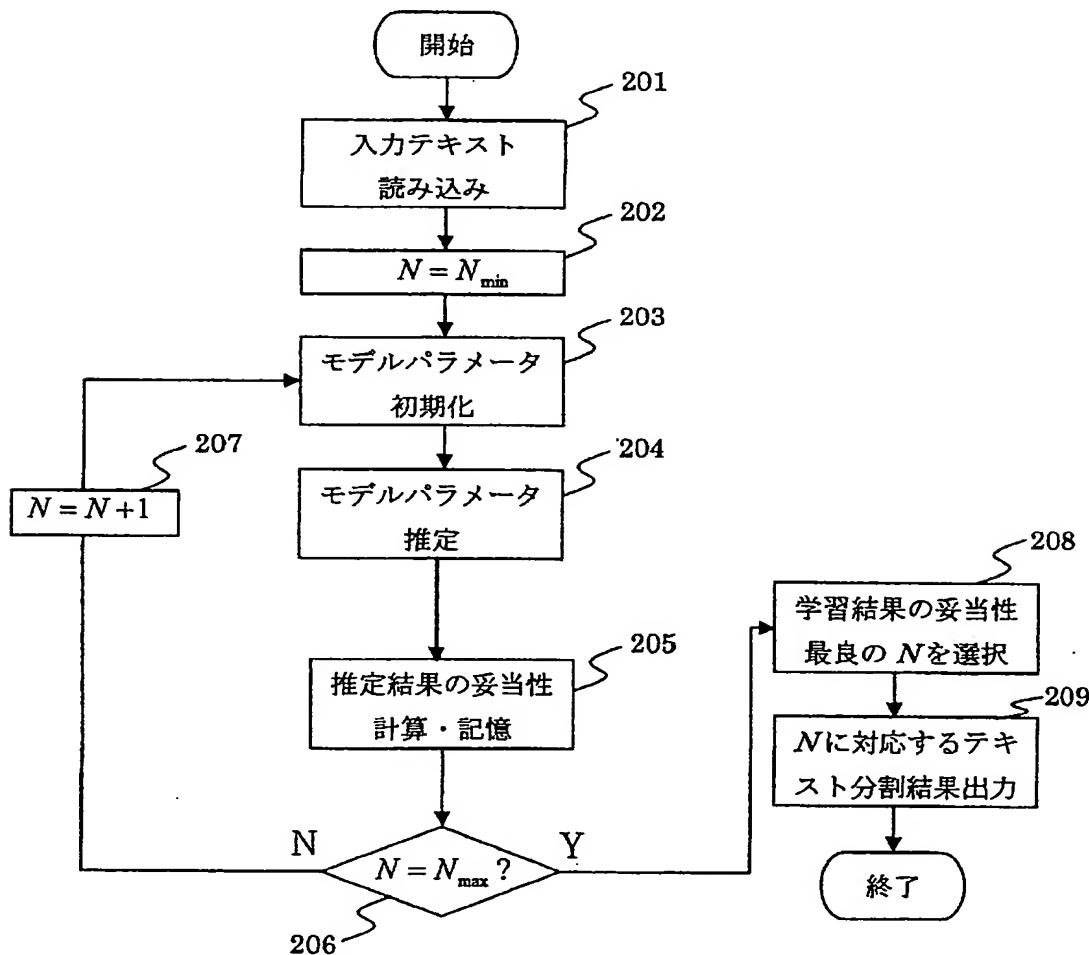
- 生成された確率モデルを規定するモデルパラメータの初期値を出力するステップと、
出力されたモデルパラメータの初期値と、処理対象のテキスト文書とともにとづいて、
このテキスト文書に応じたモデルパラメータを推定するステップと、
推定されたモデルパラメータにもとづいて、処理対象のテキスト文書を話題ごとに分
割するステップと
をコンピュータに実行させるためのプログラムを記録した記録媒体。
- [12] テキスト文書を構成する各々の単語がどの話題に属するかを隠れ変数に、各々の
単語を観測変数にそれぞれ対応付けた確率モデルを生成する仮モデル生成手段と、
前記仮モデル生成手段によって生成された確率モデルを規定するモデルパラメー
タの初期値を出力するモデルパラメータ初期化手段と、
前記モデルパラメータ初期化手段から出力されたモデルパラメータの初期値と、処
理対象のテキスト文書とともにとづいて、このテキスト文書に応じたモデルパラメータを
推定するモデルパラメータ推定手段と、
前記モデルパラメータ推定手段によって推定されたモデルパラメータにもとづいて
、処理対象のテキスト文書を話題ごとに分割するテキスト分割結果出力手段と
を備えることを特徴とするテキスト処理装置。
- [13] 請求項12に記載のテキスト処理装置において、
前記仮モデル生成手段は、複数の確率モデルを生成する手段を備え、
前記モデルパラメータ初期化手段は、複数の確率モデルのそれぞれのモデルパラ
メータの初期値を出力する手段を備え、
前記モデルパラメータ推定手段は、複数の確率モデルのそれぞれのモデルパラメ
ータを推定する手段を備え、
さらに、前記モデルパラメータ推定手段によって推定された複数のモデルパラメ
ータにもとづいて、複数の確率モデルから1つの確率モデルを選択し、前記テキスト分
割結果出力手段に対して、当該確率モデルについて処理を行わせるモデル選択手
段を備えることを特徴とするテキスト処理装置。

- [14] 請求項12に記載のテキスト処理装置において、
確率モデルは、隠れマルコフモデルであることを特徴とするテキスト処理装置。
- [15] 請求項14に記載のテキスト処理装置において、
隠れマルコフモデルは、一方向型の構造を有することを特徴とするテキスト処理装置。
- [16] 請求項14に記載のテキスト処理装置において、
隠れマルコフモデルは、離散出力型であることを特徴とするテキスト処理装置。
- [17] 請求項12に記載のテキスト処理装置において、
前記モデルパラメータ推定手段は、最尤推定および最大事後確率推定のいずれかを用いてモデルパラメータを推定する手段を備えることを特徴とするテキスト処理装置。
- [18] 請求項12に記載のテキスト処理装置において、
前記モデルパラメータ初期化手段は、モデルパラメータを確率変数とする分布を仮定し、この分布を規定するメタパラメータの初期値を出力する手段を備え、
前記モデルパラメータ推定手段は、出力されたメタパラメータの初期値と、処理対象のテキスト文書とともにとづいて、このテキスト文書に応じたメタパラメータを推定する手段を備えることを特徴とするテキスト処理装置。
- [19] 請求項18に記載のテキスト処理装置において、
前記モデルパラメータ推定手段は、ベイズ推定を用いてメタパラメータを推定する手段を備えることを特徴とするテキスト処理装置。
- [20] 請求項13に記載のテキスト処理装置において、
前記モデル選択手段は、赤池情報量基準、最小記述長基準およびベイズ事後確率のいずれかを用いて確率モデルを選択する手段を備えることを特徴とするテキスト処理装置。

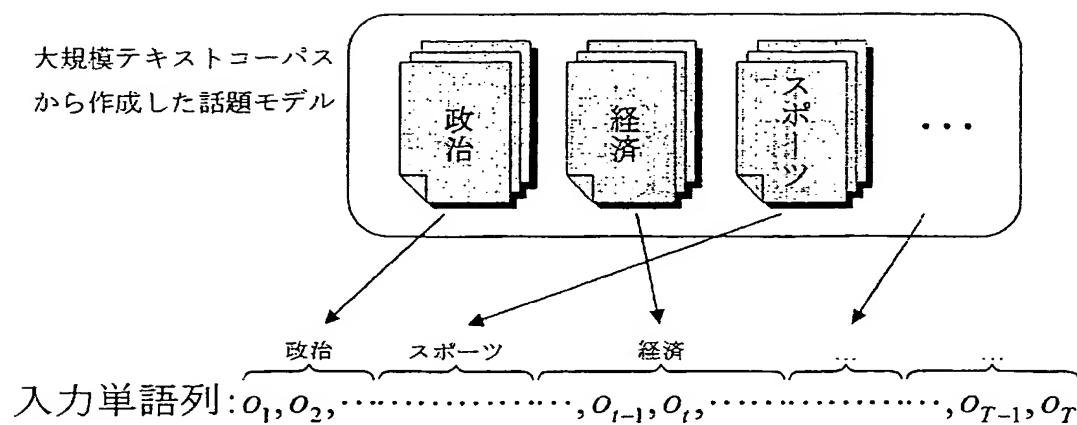
[図1]



[図2]



[図6]



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

BLACK BORDERS

IMAGE CUT OFF AT TOP, BOTTOM OR SIDES

FADED TEXT OR DRAWING

BLURRED OR ILLEGIBLE TEXT OR DRAWING

SKEWED/SLANTED IMAGES

COLOR OR BLACK AND WHITE PHOTOGRAPHS

GRAY SCALE DOCUMENTS

LINES OR MARKS ON ORIGINAL DOCUMENT

REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY

OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.